Conceptos Básicos de Colas

CONTENIDOS

1. Introducción
2. Elementos básicos de un modelo de colas
3. Notación Kendall: A / B / c / k / m / Z
4. Medidas de comportamiento. Sistema estable y sistema saturado
5. Ecuaciones de coste y fórmula(s) de Little
6. Comportamiento de transición y estacionario. Probabilidades límites. Propiedad PASTA

4. Medidas de rendimiento del sistema

Serán útiles tanto para el propio cliente que ingresa en el sistema como para el planificador, gestor o analista del mismo.

Como estos modelos representan sistemas dinámicos, los valores de estas medidas varían con el tiempo. Sin embargo, analizaremos los resultados que se obtienen cuando el sistema está en **equilibrio**, es decir, el comportamiento transitorio ha finalizado, está en estado estacionario, el sistema se ha normalizado y los valores de las medidas de comportamiento son independientes del tiempo.

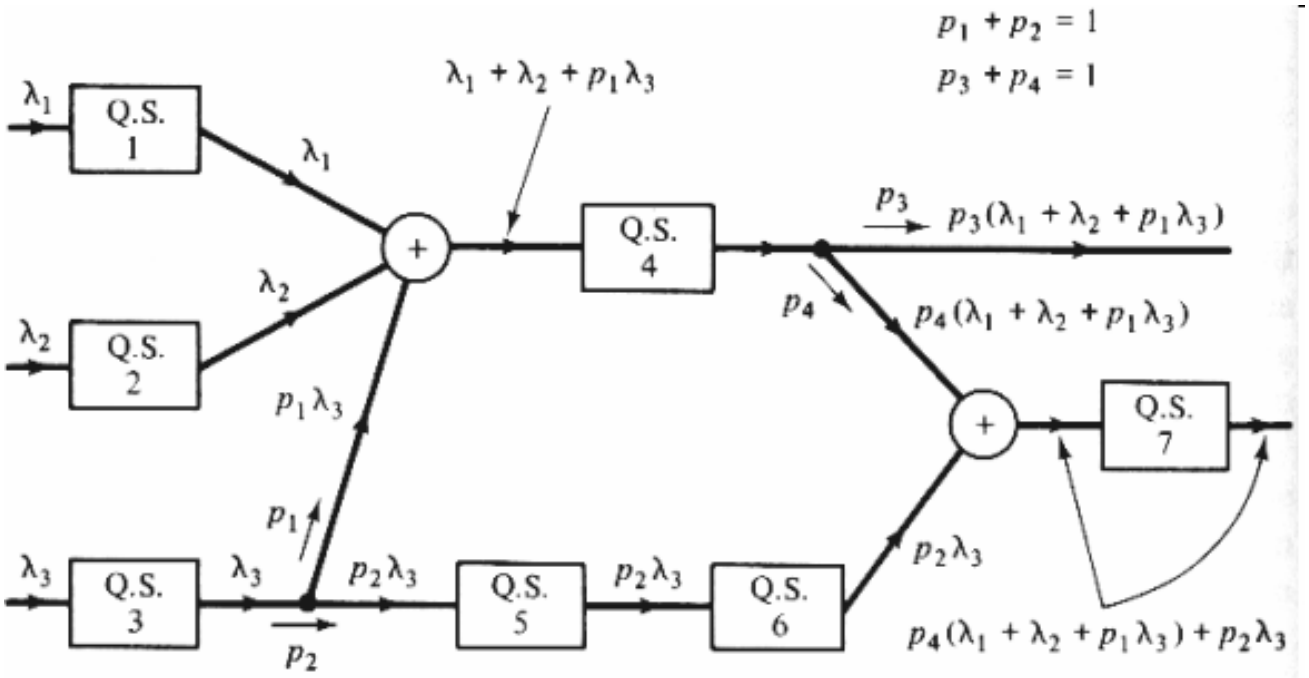
Entonces, se verifica que la tasa a la que los clientes llegan al sistema es igual a la tasa a la que salen del sistema. A este sistema también se le denomina sistema estable.

Las soluciones transitorias sólo están disponibles en forma cerrada para sistemas muy simples, y en casos más generales, habrá que recurrir a técnicas de cadenas de Markov (*Procesos Estocásticos*).

4. Medidas de rendimiento del sistema

Las medidas más importantes del rendimiento de un sistema de colas son:

1. **Probabilidad de que haya *n* clientes en el sistema**
2. **Trabajo**
3. **Intensidad de tráfico**
4. **Utilización o uso del servidor**
5. **Productividad del sistema**
6. **Tiempos de permanencia en el sistema y en la cola**
7. **Número medio de clientes en el sistema y en la cola**



4. Medidas de rendimiento del sistema

1. Probabilidad de que haya *n* clientes en el sistema

Los valores medios de muchas otras medidas podrán deducirse de ella:

$$\pi_n = P \text{ (hay } n \text{ clientes en el sistema)} = \lim_{t\to\infty} p_n(t)$$

Por ejemplo, $\pi_0 = 0.4$ indica que a largo plazo el sistema estará vacío el 40% del tiempo.

2. Trabajo

La llegada de cada cliente supone al sistema una cantidad media de trabajo, que suele medirse en tiempo y que coincide con su tiempo medio de servicio $W_s = 1/\mu$.

3. Intensidad de tráfico

Se define como:

$$r = \lambda/\mu = \lambda\, E(s) = E(s) / E(T)$$

Si λ es el número medio de llegadas por unidad de tiempo, la cantidad total de trabajo que tiene el sistema en media por unidad de tiempo es $r = \lambda\, W_s = \lambda/\mu$.

4. Medidas de rendimiento del sistema

Por ejemplo, si el tiempo entre llegadas fuese siempre constante e igual a 30 segundos y el de servicio 15 segundos, entonces el servidor estaría ocupado la mitad del tiempo: $r = 15/30 = 0.5$.
Si este servidor fuese reemplazado por otro más lento, que emplea 45 segundos en dar servicio, entonces $r = 45/30 = 1.5$. Es decir, necesitaría dar 45 segundos de servicio cada 30 segundos, lo que es imposible, a no ser que se añadiese otro servidor.

Es, por tanto, una medida del número mínimo de canales que se necesitan para atender el flujo de clientes que llegan y que hace que el sistema sea estable.

Por ejemplo, si $\lambda = 9$ clientes por día y $\mu = 2$ clientes por día, entonces $r = 4.5$ y necesitaríamos al menos 5 canales para poder satisfacer las necesidades de los clientes.

Esto es, el número de canales requeridos es el menor entero positivo c tal que $r / c < 1$. Conocido r, el gestor tiene las opciones de aumentar el número de canales si no son suficientes o bien aumentar su velocidad de servicio (es decir, μ) para disminuir r.

4. Medidas de rendimiento del sistema

4. Utilización o uso del servidor

El uso o (factor de) utilización de un servidor si hay varios (c) en el sistema, es la fracción media de servidores activos (proporción media de tiempo que cada servidor está ocupado).

Por tanto, como $c\mu$ es la tasa global de servicio, suponiendo que el tráfico está igualmente repartido entre todos los servidores,

$$p = r / c = \lambda / c\mu$$

representa también la cantidad media de trabajo que recibe cada servidor.

Si sólo hay un servidor, p es la proporción de tiempo que está ocupado, es decir, $p = r$, siempre que no haya límite sobre la capacidad del sistema.

p es una medida de la congestión del sistema y puede usarse para formular la condición de comportamiento estable mencionada antes, $p < 1$ → para que el sistema soporte el nivel de demanda, en media debe ser menor el número de clientes que llegan en una unidad de tiempo que el número de clientes que pueden ser atendidos.

4. Medidas de rendimiento del sistema

Si no, el número de clientes almacenados en la cola crecerá sin límite con el paso del tiempo. Por eso llamaremos situación de congestión a $p \geq 1$.

La situación de igualdad, $p = 1$, da lugar a congestión salvo en contadas excepciones, como en una cola $D/D/c$ con tráfico no aleatorio.

5. Productividad del sistema

La productividad del sistema o caudal o paso a través del sistema, $\boldsymbol{\lambda}$, es el número medio de clientes cuyo servicio se completa en una unidad de tiempo, es decir, es la tasa de salida del sistema.

Sistema con capacidad ilimitada, $\boldsymbol{\lambda} = \min\{\lambda, c\mu\}$.
Sistema congestionado, $\boldsymbol{\lambda} = c\mu$.
Sistema estable y sin pérdidas, como la tasa de salida coincide con la tasa de llegadas, se tiene que $\boldsymbol{\lambda} = \lambda = \rho c\mu$

Recordemos de nuevo que si la capacidad del sistema es finita, la productividad puede ser diferente a la tasa externa de llegadas y será $\boldsymbol{\lambda} < \min\{\lambda, c\mu\}$.

4. Medidas de rendimiento del sistema

6. Tiempos de permanencia en el sistema (w) y en la cola (q)

Las medias de dichos tiempos las denotaremos como $W = E(w)$ y $W_q = E(q)$, respectivamente. Por tanto, $w = q+s$ y tomando esperanzas, $W = W_q+E(s) = W_q+1/\mu = W_q+ W_s$.

A veces utilizaremos las funciones de distribución de estas tres variables aleatorias, denotadas como $F_w(t)$, $F_q(t)$, $F_s(t)$, respectivamente.

7. Número de clientes en el sistema (N) y en la cola (N_q)

La media de la variable aleatoria N, que toma los valores 0,1,2..., será

$$L = E(N) = \sum_{n=1}^{\infty} n\pi_n$$

Podemos decir que $N_q = \max\{0, N-c\}$. Su esperanza matemática la denotaremos con $L_q = E(N_q)$.

Si N_s indica el número de clientes siendo servidos, con media L_s, entonces $N = N_q + N_s$ y $L = L_q + L_s$.

5. Ecuaciones de coste y fórmulas de Little

Fórmulas de Little:

$$L = \lambda W$$
$$L_q = \lambda W_q$$
$$L_s = \lambda W_s$$

El teorema de Little es válido bajo condiciones muy generales de un sistema estable, cualquier número de servidores y para todas las disciplinas de colas.

Intuitivamente puede explicarse de la siguiente forma:

Un cliente que acaba de llegar saldrá del sistema, en promedio, después de un tiempo W.

Cuando salga, quedarán en el sistema L clientes en promedio.

Cada uno de estos clientes ha llegado, en promedio, tras un tiempo $1/\lambda$.

El tiempo que han tardado en llegar estos L clientes es $L \times (1/\lambda)$ y ha de ser igual al tiempo que nuestro cliente ha pasado en el sistema, W.

El razonamiento es análogo para la cola y el servicio.

5. Ecuaciones de coste y fórmulas de Little

Otra explicación se basa en la idea de que los clientes han de pagar por estar en el sistema. En este caso, esperamos que se cumpla

$$\left(\begin{array}{c}\text{Tasa media de}\\ \text{ingresos del sistema}\end{array}\right) = \lambda \times \left(\begin{array}{c}\text{Cantidad media}\\ \text{pagada por cliente}\end{array}\right)$$

Entonces, si cada cliente paga 1 euro por cada unidad de tiempo que pasa en el sistema, la igualdad anterior se convierte en $L = \lambda W$; si paga 1 euro por cada unidad de tiempo que pasa en la cola, se obtiene $L_q = \lambda W_q$ y si paga 1 euro por cada unidad de tiempo que pasa en el servidor, se verifica $L_s = \lambda W_s$

Si multiplicamos por λ la igualdad $W = W_q + E(s)$, obtenemos $\lambda W = \lambda W_q + \lambda E(s)$. Por las fórmulas de Little, resulta

$$L = L_q + r$$

que sirve para cualquier modelo $G/G/c$ y nos dice que el número medio de clientes en el sistema es el número medio de clientes en cola más el número medio de clientes en los servidores.

6. Comportamiento de transición y estacionario. Probabilidades límites. Propiedad PASTA

Nótese que siempre asumimos que el sistema ha alcanzado el equilibrio, pues si estuviera en estado transitorio habría que considerar la dependencia del tiempo que el sistema lleva funcionando y de las condiciones iniciales.

Hay dos variantes de las probabilidades π_n:

- proporción a_n ($n \geq 0$) de clientes que encuentran n en el sistema al llegar
- proporción b_n ($n \geq 0$) de clientes que dejan n en el sistema al salir.

a_n corresponde a lo que una llegada observa,
b_n a lo que una salida observa y
π_n a lo que observaría alguien desde fuera.

Las tres probabilidades no tienen por qué coincidir.

Proposición. En un sistema en el que los clientes llegan de uno en uno y se sirven de uno en uno, $a_n = b_n, \square n$.

6. Comportamiento de transición y estacionario. Probabilidades límites. Propiedad PASTA

Demostración. Cuando el sistema pasa de tener n clientes a n+1 es cuando una llegada ve n clientes en el sistema. De igual modo, cuando pasa de n +1 a n es cuando una salida deja n en el sistema.

En cualquier intervalo de tiempo, el número de transiciones de n a n +1 se diferencia en 1 del número de transiciones de n +1 a n.

Por tanto, las tasas de transiciones de n a n +1 y de n +1 a n coinciden, o equivalentemente, también lo hacen la tasa a la que las llegadas encuentran n clientes y la tasa a la que las salidas dejan n clientes.
Se llega así al resultado deseado porque las tasas de llegadas y salidas globales deben coincidir.

En media, las llegadas y salidas ven siempre el mismo número de clientes, aunque en general no ven tiempos medios. Hace falta que las llegadas sean de Poisson para que coincidan las tres probabilidades a_n, b_n y π_n.

6. Comportamiento de transición y estacionario. Probabilidades límites. Propiedad PASTA

Esta **propiedad** se denomina **PASTA** (Poisson Arrivals See Time Averages): Las llegadas de Poisson siempre ven tiempos medios, es decir, $\pi_n = a_n, \forall n$.

Demostración. Si $A(t, t+\Delta t)$ denota la llegada de un cliente en el intervalo $(t, t+\Delta t)$, se tiene

$$\begin{aligned} a_n(t) &= \lim_{\Delta t \to 0} P(N(t) = n \mid A(t, t+\Delta t)) \\ &= \lim_{\Delta t \to 0} \frac{P(A(t, t+\Delta t) \mid N(t) = n) P(N(t) = n)}{P(A(t, t+\Delta t))} \\ &= \lim_{\Delta t \to 0} \frac{P(A(t, t+\Delta t)) P(N(t) = n)}{P(A(t, t+\Delta t))} = P(N(t) = n) = p_n(t). \end{aligned}$$

La tercera igualdad se debe a la propiedad de pérdida de memoria de la distribución exponencial, al ser las llegadas de Poisson.
Hemos obtenido que la distribución de lo que una llegada en el tiempo t observa es la misma que la distribución del estado del sistema en el tiempo t. De ahí, a largo plazo,

$$a_n = \lim_{t \to \infty} a_n(t) = \lim_{t \to \infty} p_n(t) = \pi_n.$$

6. Comportamiento de transición y estacionario. Probabilidades límites. Propiedad PASTA

Por tanto, la distribución de probabilidad que observarán los clientes al llegar será la distribución de probabilidad a lo largo del tiempo.

La tarea del analista de sistemas de colas consiste en caracterizar adecuadamente el modelo que va a utilizar para analizar el sistema real bajo estudio y todas las variables implicadas, pero también los costes asociados.

Una vez realizado este estudio, si el analista pudiera encontrar una relación a optimizar, de acuerdo a algún criterio, llegaría a determinar un sistema óptimo.

Sin embargo, no siempre es fácil llegar a este tipo de relación. En sistemas complejos nos deberemos conformar con formular diversas soluciones alternativas y obtener cuál de ellas es la mejor según ciertos criterios prefijados.

Cuando el problema no pueda resolverse por medios analíticos, recurriremos a métodos numéricos o a la simulación.